Int. J. Dev. Biol. 60: 141-150 (2016)
doi: 10.1387/ijdb.160108jb

THE INTERNATIONAL JOURNAL OF DEVELOPMENTAL BIOLOGY
www.intjdevbiol.com

Genome-wide identification of enhancer elements

SARAH TULIN[1,#], JULIUS C. BARSI*[,2,#], CARLO BOCCONCELLI[1,#] and JOEL SMITH[1]

[1]Eugene Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, MA and [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA

ABSTRACT We present a prospective genome-wide regulatory element database for the sea urchin embryo and the modified chromosome capture-related methodology used to create it. The method we developed is termed GRIP-seq for genome-wide regulatory element immunoprecipitation and combines features of chromosome conformation capture, chromatin immunoprecipitation, and paired-end next-generation sequencing with molecular steps that enrich for active *cis*-regulatory elements associated with basal transcriptional machinery. The first GRIP-seq database, available to the community, comes from *S. purpuratus* 24 hpf embryos and takes advantage of the extremely well-characterized *cis*-regulatory elements in this system for validation. In addition, using the GRIP-seq database, we identify and experimentally validate a novel, intronic *cis*-regulatory element at the *onecut* locus. We find GRIP-seq signal sensitively identifies active *cis*-regulatory elements with a high signal-to-noise ratio for both distal and intronic elements. This promising GRIP-seq protocol has the potential to address a rate-limiting step in resolving comprehensive, predictive network models in all systems.

KEY WORDS: *GRIP-seq, chromatin conformation capture, anti-Pol-II, sea urchin development*

Introduction

In any given nucleus, individual genes are expressed or not expressed due to the binding of transcription factors to specific regulatory elements. Regulatory elements may be located at a distance from the transcription start site (TSS), either upstream or downstream, including in the transcribed regions of the gene. The fundamental relationships between regulatory elements and TSSs control the differential expression of genes that provides the basis for differences in cell function. Identifying regulatory elements and their cognate TSSs is thus an important research goal in many areas related to life science, human health, and disease. Despite a number of available methods for identifying regulatory elements, mapping and characterizing every regulatory region of a genome and validating their relationship with a TSS remains a serious challenge. In particular, determining gene network models at a level of resolution that allows the researcher to make specific predictions of which interactions between a transcription factor and a regulatory element will drive alternate outcomes requires detailed knowledge of many of these relationships often down to the base pair level. This level of regulatory analysis is a rate-limiting step in network modeling. We present here a time-saving method to putatively map an organism's genomic regulatory architecture and to identify elements at high resolution that are actively involved in differential gene expression. Specifically, the aims of this method are (1) to map putative regulatory elements genome-wide, (2) to be applicable to any model system, (3) to be equally sensitive to sites involved in activation and repression, (4) to identify elements located distally and/or within introns, and (5) to link regulatory regions to their cognate core promoters/TSSs. This method is termed GRIP-seq for genome-wide regulatory element immunoprecipitation followed by next-generation sequencing.

Abbreviations used in this paper: 3C, chromosome conformation capture; 4C, circularized chromosome conformation capture; 5C, chromosome conformation capture carbon copy; ChIP, chromatin Immunoprecipitation; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; CDS, coding DNA sequence; CRE, *cis*-regulatory element; CRM, *cis*-regulatory module; dGRN, developmental gene regulatory network; Hi-C, a method extending 3C with massively parallel sequencing; GRIP-seq, genome-wide regulatory immunoprecipitation paired with next generation sequencing; MACS, Model based Analysis for ChIP-Seq; Pol II, RNA polymerase; TSS, transcription start site.

***Address correspondence to:** Julius C. Barsi. Division of Biology and Biological Engineering, California Institute of Technology, 1200 East California Boulevard, Mail Code 156-29, Pasadena, CA 91125, USA. Tel: +54-929-4469-0388. E-mail: barsi@caltech.edu - Web: http://www.its.caltech.edu/~barsi/

 # Note: These authors contributed equally.

Supplementary Material (four figures) for this paper is available at: http://dx.doi.org/10.1387/ijdb.160108jb

Accepted: 14 June 2016.

ISSN: Online 1696-3547, Print 0214-6282

GRIP-seq, like other chromosome capture methods, takes advantage of the looped DNA conformation created during the activation and repression of gene transcription. Transcription factors bound to regulatory elements stabilize a bend in the DNA that enables them to physically contact the basal promoter region (Levine *et al.*, 2014; Smith, Davidson, 2009). This physical linkage brings the regulatory elements responsible for transcription (or repression of transcription) in close proximity to the promoter and bound RNA Polymerase II (Pol II). Experimentally, the complex consisting of an enhancer, promoter, and Pol II can be fixed by formaldehyde cross-linking and captured with chromatin immunoprecipitation (ChIP). The captured DNA can then be fully sequenced and mapped back to the genome. A first step towards exploiting this "looping principle" to target active gene regulation came with the development of the chromosome conformation capture (3C) technique (Dekker *et al.*, 2002), where enzymatic DNA cleavage and PCR analysis are used to reveal a regulatory interaction at a particular locus. A successive variation, 4C, focusing on chromosome interactions in *trans,* uses inverse PCR to map areas of the genome associated with a single locus (Zhao *et al.*, 2006). Expanding 4C capabilities, 5C, provides an outline of the interactions between two DNA regions of interest by creating selective PCR primers based on the sequences of those known regions (Dostie *et al.*, 2006). With the advent of next-generation sequencing, new technologies based on 3C methodology have emerged. The Hi-C method produces mega-base scale genome-wide maps of three-dimensional interactions by selectively purifying only proximity-ligated DNA (as would occur due to DNA looping) (van Berkum *et al.*, 2010). Hi-C shows evidence for the existence of chromatin neighborhoods, but the method is limited in the ability to assess specific *cis*-regulatory elements at a resolution below megabase scale. Another method to make use of 3C technology for the purpose of identifying long-range chromatin interactions is termed ChIA-PET (Fullwood *et al.*, 2009). ChIA-PET is a powerful technique, but the lengthy, elaborate protocol can require a large amount of starting material, is costly and is not trivial to replicate (Daniel *et al.*, 2014; Li *et al.*, 2014). Consequently, ChIA-PET is frequently used on cell lines where large quantities of cells can be harvested. Lastly, when used in conjunction with an antibody that recognizes only a single transcription factor, as many published ChIA-PET experiments have done, only a small subset of interactions occurring in the genome can be identified at a given time.

GRIP-seq combines elements from both chromatin conformation capture and ChIA-PET, with additional molecular steps that enrich for genomic fragments cross-linked to the TSS. From the outset, GRIP-seq has been designed to optimize conditions favoring less starting material, enhancing the signal-to-noise ratio at distal (as well as intronic) regulatory elements, and ultimately, determining the gene(s) that a particular regulatory element governs. GRIP-seq maps interactions between promoters and putative regulatory elements on a genome-wide scale. Importantly, it relies on steps that accommodate all animal model systems in use today. In the present report, we present our GRIP-seq database for the 24-hr sea urchin embryo which is publically available to the community and evaluate GRIP-seq signal across multiple genes for which experimentally validated regulatory elements are known, including *blimp1* and *foxa*. We demonstrate that GRIP-seq accurately identifies the genomic coordinates of these regulatory elements. We extend this analysis by utilizing GRIP-seq data to identify putative regulatory elements at the *onecut* locus, which we then experimentally corroborate. Lastly, GRIP-seq signal from *nodal* and *onecut* loci demonstrate the technique's potential to capture the physical contact between active *cis*-regulatory elements and their cognate core promoter.

A limitation of methods relying on ChIP-seq technology is that they often over-state the ability to infer causal functionality from protein-chromatin binding. We thus approached our method verification with known functionality as an absolute requirement for test cases. Deliberately using the sea urchin embryo, which has many rigorously tested *cis*-regulatory elements, we aimed to test GRIP-seq signals for specificity and sensitivity using regulatory sites with strongly functional consequences for embryonic development. These "hub" genes confer critical network functions and we found GRIP-seq signal highest at the site of these verified regulatory regions.

Results

GRIP-seq methodology overview

A flowchart depicting an overview of the key steps in the GRIP-seq method is presented in Fig. 1. Briefly, following fixation and shearing of the cross-linked chromatin, target regulatory region-promoter complexes are captured by immunoprecipitation with an anti-Pol II antibody. The isolated chromatin fragments undergo ligation conditions while still on the immunoprecipitation bead, to enrich for close-proximity ligation between the fixed TSS-containing fragments and specific regulatory element-containing fragments, ultimately forming a circularized product. Non-circular DNA fragments are then digested with a plasmid-safe exonuclease enzyme. Circular ligation products are further enriched using rolling circle amplification (Dean *et al.*, 2001; Hutchison *et al.*, 2005). A sequencing library is then prepared with the amplified products for paired-end sequencing. Mapping the resulting paired-end sequence reads onto the genome results in a genome-wide map of all putative regulatory elements and TSSs that are active at the time of the assay and relevant for resolving the transcriptional network in question. Further peak calling analysis was performed using MACS software and the called peaks provide additional information for planning further experimental validation experiments (Feng *et al.*, 2012).

The protocol steps are described in detail within the Materials and Methods section. All of the essential steps (delineated in Fig. 1) were chosen to ensure compatibility with known animal model systems. Our pioneer dataset comes from the *S. purpuratus* sea urchin embryo which has the benefit of a well-characterized developmental gene regulatory network (dGRN) and base pair resolution of many validated *cis*-regulatory elements which directly bind specific transcription factor drivers (Barsi, Davidson, 2016; de-Leon, Davidson, 2010; Nam *et al.*, 2007; Smith *et al.*, 2008). We leverage this knowledge of bona fide *cis*-regulatory elements against our GRIP-seq database in order to validate it.

GRIP-seq identifies all experimentally determined distal regulatory elements controlling **blimp1b** *expression*

In *S. purpuratas* embryos, the spatiotemporal expression of *blimp1b* is controlled by an upstream distal regulatory region. In addition, a proximal element and an intronic element have also been shown to contribute to amplifying expression (Smith *et al.*, 2007; Smith *et al.*, 2008). In the distal upstream element, termed

CR2 in previous reports and in Fig. 2, two Blimp1 binding sites and overlapping twin TCF binding sites lying approximately 8 kb upstream from the start of transcription are responsible for shaping the correct spatial and temporal expression of Blimp1b. As shown in Fig. 2, GRIP-seq signal is observed in the *blimp1b* regulatory regions both at the start of transcription and within the region previously identified as CR2, containing the twin Blimp1/TCF binding sites (Fig. 2A). Given the importance of this particular element, it was dissected in further detail than other CREs and serves as an example of the high specificity and sensitivity of GRIP-seq. The GRIP-seq data further reveal a shoulder profile immediately proximal to CR2; this element is known to contain amplification activity though not spatiotemporal determinants. Lastly, GRIP-seq signal at the *blimp1b* locus further identifies a region in the first intron denoted CR6 and this element is also known to contribute to the amplitude of expression but not spatial nor temporal restriction. Thus, GRIP-seq sensitively and specifically identifies the most important elements driving *blimp1b* expression, an example of a locus which has been interrogated particularly well. The strong agreement between the GRIP-seq signal located outside of the coding sequence (CDS) and that of previously characterized *cis*-regulatory elements (with a surprisingly clear signal recovered from the distal regulatory element), demonstrate the capacity of GRIP-seq for reducing the search space for prospective *cis*-regulatory analysis.

GRIP-seq identifies the full complement of distal and proximal activation and repressor modules at the foxa *locus*

GRIP-seq signal at the *foxa* locus correlates with validated *cis*-regulatory modules (CRMs) identified in de-Leon and Davidson, 2010 (de-Leon, Davidson, 2010). Four CRMs within the *foxa* region have been published, designated "F", "I", "J", and "K." Our dataset shows discernable signal in all of these modules (Fig. 2B). Module F is located about 10 kb upstream from the TSS and contains repressor elements. GRIP-seq signal within module F demonstrates that GRIP-seq can detect regulatory interactions that are repressive in nature in addition to those that correspond to activation (similar to CR2 at the *blimp1b* locus, described above, which mediates both repression and activation). GRIP-seq signal

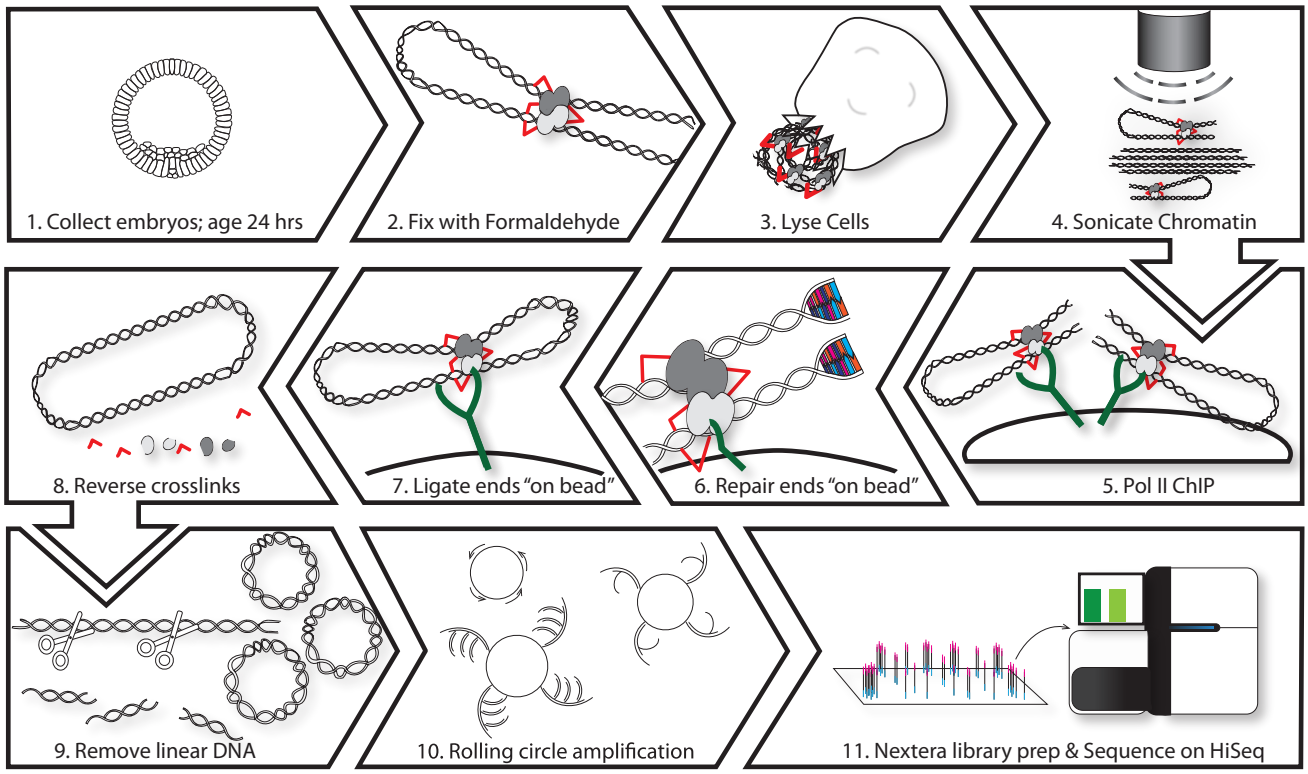


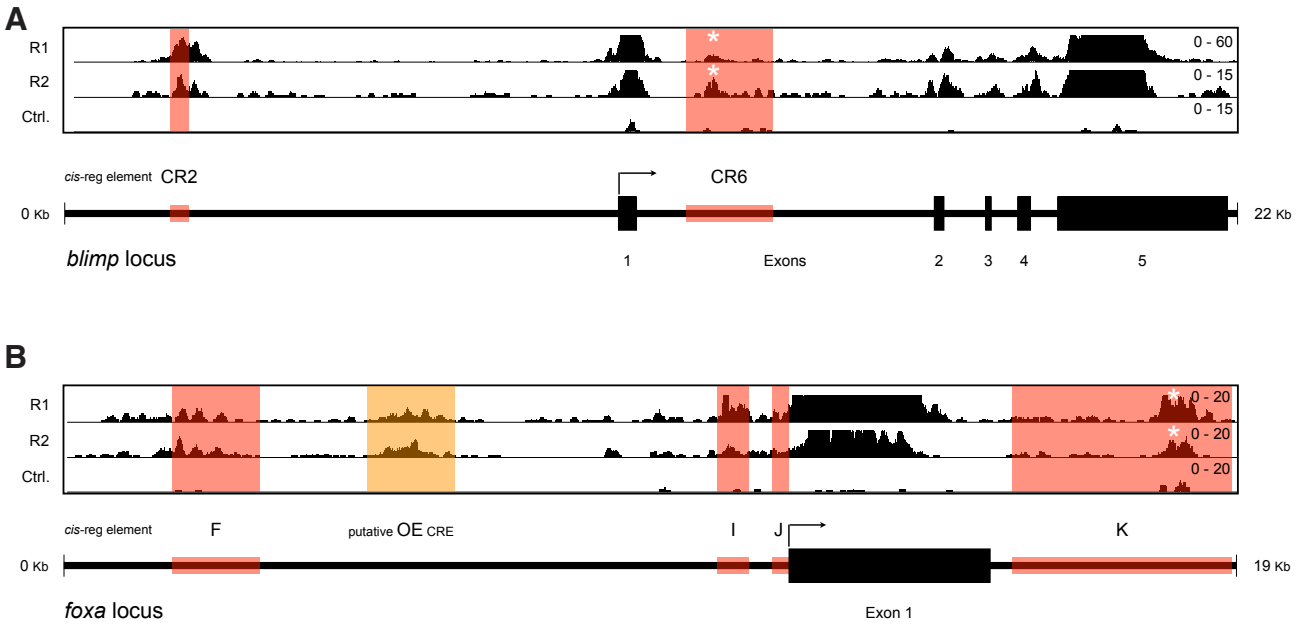
Fig. 1. GRIP-Seq protocol overview flowchart depicts important steps. *Highlights of the GRIP-seq protocol depicting the universal nature of the method and the major steps that enrich for active regulatory elements.* **(1)** *Embryos are collected and aged.* **(2)** *Embryos are fixed with formaldehyde to crosslink (red links) protein-protein and protein-DNA interactions. In this illustration, a transcription factor (dark grey shape) has bound Pol II (light grey shape), and formaldehyde crosslinks have formed to stabilize the 3D conformation of the looped DNA.* **(3)** *Cells are lysed, releasing crosslinked chromatin.* **(4)** *The chromatin is sheared by sonication.* **(5)** *Chromatin is immunoprecipitated with beads coupled to anti-Pol II antibodies, which isolates DNA-protein complexes fixed to Pol II protein.* **(6)** *The DNA is end-repaired while still attached to the beads, resulting in blunt end overhangs.* **(7)** *The repaired DNA ends are ligated together while still attached to the bead, favoring close-proximity ligation products. Circularized DNA products are formed, some of which include both sequence from the TSS region and sequence from a regulatory element region.* **(8)** *Protein-protein and protein-DNA crosslinks are reversed with heat and proteinase K.* **(9)** *Any remaining linear DNA fragments that were not ligated are digested with a plasmid-safe exonuclease enzyme.* **(10)** *Circular ligation products are then further enriched using rolling circle amplification and* **(11)** *prepared into a sequencing library for next generation sequencing. Mapping the sequence pairs results in a genome-wide map of all putative regulatory elements and transcription start sites. The entire protocol is detailed in the methods section. Due to the versatility of the reagents and methods, GRIP-seq can be readily applied to any organism or cell culture model.*

also identifies the three other tested regulatory elements (I, J, and K). Furthermore, GRIP-seq identifies a smaller, more distinct, region for module K, indicating the possibility of further narrowing down the DNA sequence responsible for this element's activity. As previously reported, the genomic region between module F and module I possess the capacity to drive reporter expression restricted to the oral ectoderm (OE), though the module was not specifically isolated (de-Leon, Davidson, 2010). The GRIP-seq data point to an element in that region which is highlighted in orange in Fig. 2B and is therefore a likely candidate for the element governing the OE expression of *foxa*, shown in Fig. 2C. Similar to what was seen at the *blimp1b* locus, this further supports the capacity to use GRIP-seq as a method for the prospective identification of verifiable *cis*-regulatory elements.

GRIP-seq identifies novel regulatory elements within the onecut *locus*

In the sea urchin embryo, *onecut* [previously referred to as *hnf6* (Otim *et al.*, 2004)] encodes a pioneer factor whose zygotic expression pattern delineates a neurogenic field from which the ciliated band will arise (Barsi, Li, *et al.*, 2015; Otim *et al.*, 2004; Poustka *et al.*, 2004). The ciliated band, once fully developed, is a structure that confers upon the larva the ability to both swim and feed. It is composed, as its name implies, of a band of differentiated cells each possessing long cilia. To our knowledge, *onecut* is the first transcription factor to be expressed throughout this embryonic territory, which, from early gastrulation onward, resembles a nearly perfect ring that bisects the ectoderm into oral and aboral domains. Therefore, spatial specification of this particular transcription factor presents a uniquely challenging regulatory problem. Understanding a likely complex regulatory apparatus promises to reveal how highly resolved spatial information is integrated within a larger network of transcriptional feedback control to govern transcription.

Fig. 3A shows GRIP-seq signal spanning across the entire *onecut* locus. It is immediately apparent that the strongest signal outside of the CDS can be observed on either side of Exon 2. When



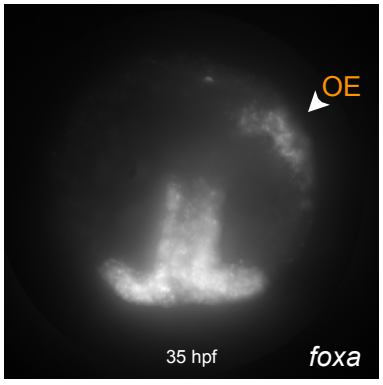


Fig. 2. GRIP-seq identifies experimentally-determined *cis*-regulatory elements as exemplified at the *blimp1b* and *foxa* loci. *GRIP-seq signal from two replicate experiments (R1 & R2) alongside their control (Ctrl.), are shown for GRIP-seq signal in the genomic regions containing* blimp **(A)** *and* foxa **(B)** *where signal is seen overlapping rigorously-tested* cis*-regulatory elements controlling their respective expression. Numbers in the upper right-hand corner represent units used for scaling purposes.* **(A)** *GRIP-seq signal is clear at the CR2* cis*-regulatory element in the* blimp *locus, which contains two Blimp1 sites and two overlapping TCF sites (indicated with light red shading on the genomic schematic and reflected on the GRIP-seq signals above). There is also GRIP-seq signal at the intronic* cis*-regulatory element CR6, which contributes to the amplitude of expression.* **(B)** *The* foxa *locus has been tested for regulatory activity and published CREs F, I, J, and K are all involved in* foxa *expression and all contain GRIP-seq signal (represented with light red shading along the genome model and overlaying the GRIP-seq signal for the corresponding locations.) Module F contains repressor elements, indicating GRIP-seq can identify elements involved in both activation and repression. GRIP-seq shows a strong signal in a smaller region for module K (light orange shading), indicating the possibility for further delineation of this element. The region from F through J was shown to have activity in the oral ectoderm (OE) region of the embryo* **(C)**, *even though F, I and J themselves do not drive expression there, indicating a potential unidentified element between F and I driving OE expression. GRIP-seq signal in that region point to a possible candidate for an additional module driving OE expression of* foxa.

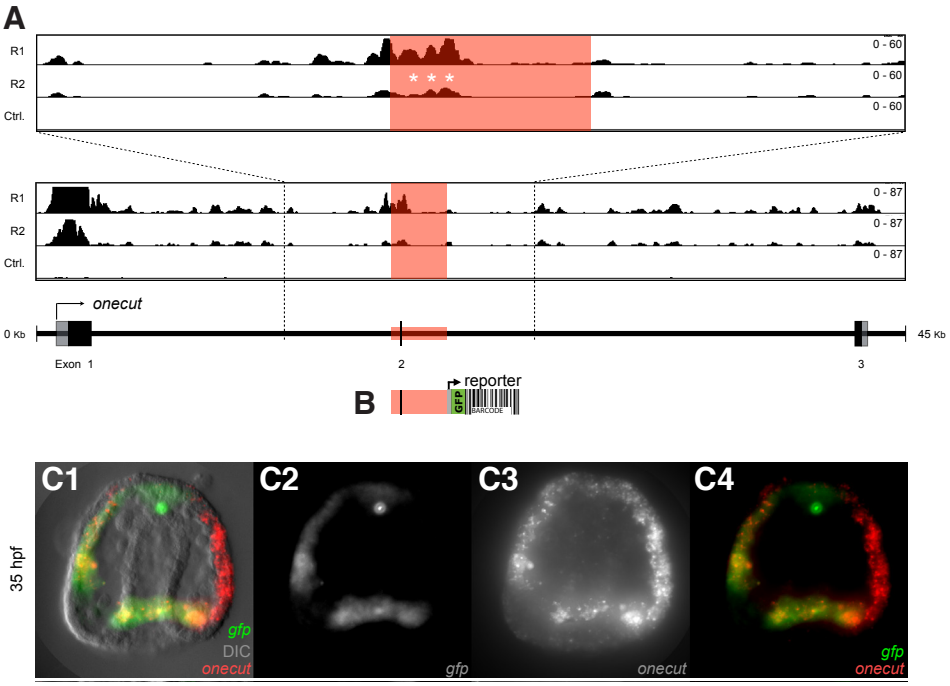


Fig. 3. Authentication of *cis*-regulatory elements identified by GRIP-seq at the *onecut* locus. *GRIP-seq signal at the* onecut *locus coincides with an experimentally-corroborated* cis*-regulatory region, active at mid-gastrula stage.* **(A)** *GRIP-seq signal from two replicate experiments (R1 & R2) alongside their control (Ctrl.), are shown for the* onecut *locus. The most prominent signal occurs at the TSS (black peak directly above Exon 1) and flanking Exon 2 (white asterisks shown within the region expanded at top). Area highlighted in light red harbors regulatory sequence; grey boxes indicate untranslated regions; solid black boxes indicate exons; and the solid horizontal line represents 45 kb of genomic sequence. Signal intensity is commensurate with peak height (numbers in the upper right-hand corner represent units used for scaling purposes).* **(B)** *A fluorescent reporter constructed of the genomic sequence highlighted in red, endogenous* onecut *basal promoter (grey box), GFP CDS (green box) and a DNA barcode (labeled as such) is capable of recapitulating the endogenous* onecut *expression pattern.* **(C)** *Double fluorescent RNA* in situ *hybridization of a 35 hpf transgenic embryo, carrying the reporter construct described in B. Endogenous* onecut *expression in red and exogenous* gfp *expression in green.* **(C1)** *Double fluorescent signal superimposed onto a DIC microphotograph of the embryo.* **(C2)** *Exogenous* gfp *mRNA (mosaic expression is attributed to the particularities of transgenesis in this model organism).* **(C3)** *Endogenous* onecut *mRNA expression pattern.* **(C4)** *Composite image generated by merging (C2,C3).*

an overlapping region (highlighted in red) is fused upstream of a reporter construct (Fig. 3B), the resulting construct proved capable of recapitulating endogenous *onecut* expression *in vivo*, both qualitatively (Figs. 3C, Supplemental Fig. S1 A-C) and quantitatively (Supplemental Fig. S1D). Many similar reporters constructed using adjacent non-coding sequence failed to drive GFP expression (data not shown). Although this particular region remains to be explored further, GRIP-seq signal points towards the sequence space most likely responsible for mediating transcriptional regulation (white asterisks).

Using GRIP-seq to screen for putative physical linkages between regulatory elements and transcription start sites

Locating and verifying a connection between any given regulatory element and the particular TSS(s) that it influences remains a time-and-resource-intensive experimental process. This information is not inferable from proximity alone and it becomes increasingly difficult to unravel connections when long distance regulators or gene-dense genomes are involved (Daniel *et al.*, 2014). Our GRIP-seq database has the potential to provide evidence linking *cis*-regulatory elements to their cognate TSSs through paired-end sequence reads where one pair maps to a TSS and the other pair to a *bona fide cis*-regulatory element. An example is shown in Fig. 4, which depicts GRIP-seq data for the sea urchin *nodal* gene. As in other deuterostomes, the regulation of *nodal* and other members of the TGFβ superfamily of signaling factors confers important early patterning information. The *cis*-regulatory control of *nodal* was previously shown to depend in part on an element in the first intron immediately upstream of Exon 2 (Fig. 4, labeled as INT) (Nam *et al.*, 2007). As shown in Fig. 4, a number of paired reads are linked between the INT element and the *nodal* TSS. The GRIP-seq signal itself for INT is lower than the high signal at the start of *nodal* Exon 2, a phenomenon we observe for most genes due to the local enrichment of Pol II. However, compared with other non-coding regions, the INT element signal is robust as is the signal showing linkage with the TSS. GRIP-seq data may be useful for making predictions genome-wide about physical connections between putative elements and basal transcription factors. Viewing the GRIP-seq data "as pairs" with a high performance visualization tool like the Broad Institute's Integrative Genomics Viewer depicts putative connections between a regulatory site and TSS, which researchers can then verify experimentally.

GRIP-seq linked pairs accurately point to a known distal regulatory element at the onecut *locus*

As we do not yet possess a map of *all* regulatory elements within the genome, it is difficult to estimate the percentage of elements recovered by GRIP-seq that exhibit linkage to a TSS. An example of a well-studied locus with GRIP-seq linked reads connecting a known regulatory element to the TSS is at the *onecut* locus. As previously mentioned, *onecut* is an interesting gene from the point of view of transcriptional regulation, as the first transcription factor whose expression delineates a unique domain in a highly specific fashion, down to the single cell level. A recently published study reveals the presence of a functional *cis*-regulatory element located approximately 40 kb away from the *onecut* TSS (Fig. 5A) (Barsi, Davidson, 2016). Remarkably, paired sequence reads connecting this element with the *onecut* TSS are recovered by GRIP-seq (Fig. 5B). No other paired reads in the vicinity were found to span such a distance. Of special note is the fact that this distal sequence

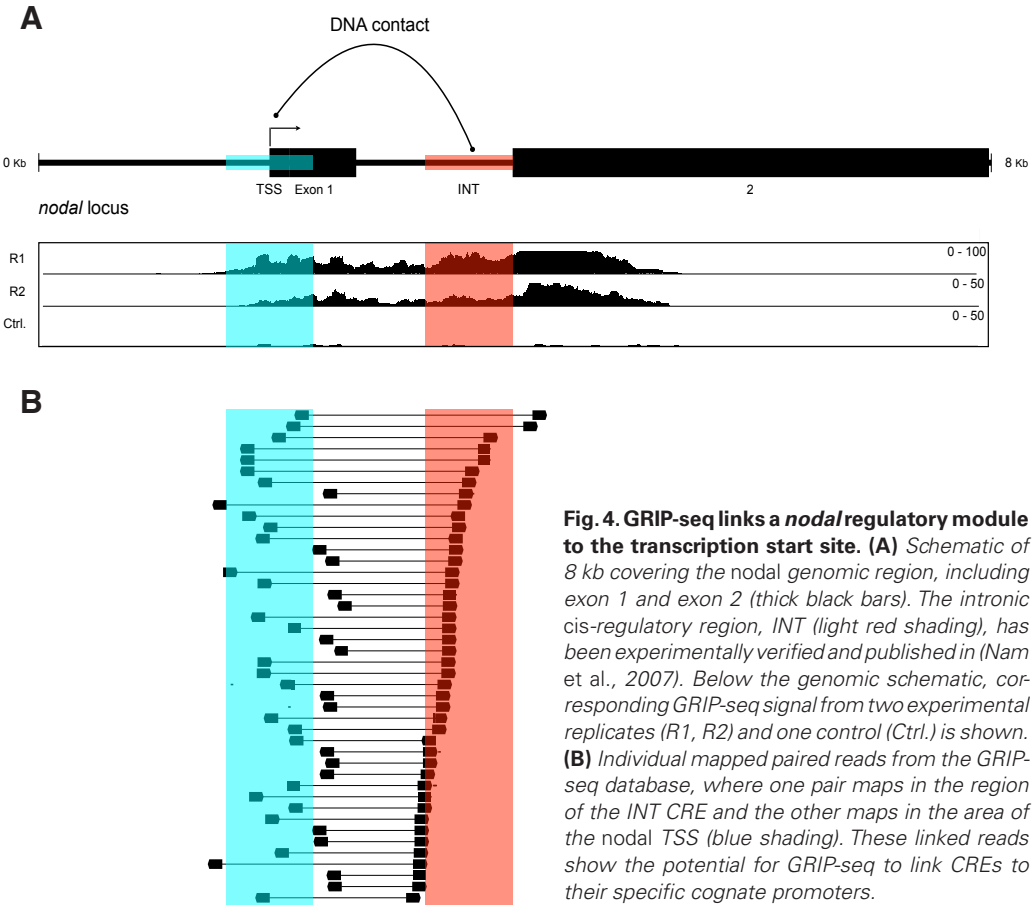


Fig. 4. GRIP-seq links a *nodal* regulatory module to the transcription start site. (A) *Schematic of 8 kb covering the* nodal *genomic region, including exon 1 and exon 2 (thick black bars). The intronic* cis*-regulatory region, INT (light red shading), has been experimentally verified and published in (Nam* et al.*, 2007). Below the genomic schematic, corresponding GRIP-seq signal from two experimental replicates (R1, R2) and one control (Ctrl.) is shown.* **(B)** *Individual mapped paired reads from the GRIP-seq database, where one pair maps in the region of the INT CRE and the other maps in the area of the* nodal *TSS (blue shading). These linked reads show the potential for GRIP-seq to link CREs to their specific cognate promoters.*

constitutes a repressive element. Published and ongoing studies show that Gsc-mediated repression is conveyed through this distal element, thereby functionally excluding *onecut* expression from the oral face. This function has been demonstrated by fusing the distal element upstream of a fluorescent reporter (Fig. 5F) and comparing its expression to an otherwise identical construct with mutated Gsc consensus binding sites (illustrated in Fig. 5, C vs. D and Supplemental Fig. S2, C-D vs. E-F). GRIP-seq paired-end analysis therefore accurately reflects a linkage between a functionally relevant distal element and the TSS through which it operates.

Discussion

We present the method and the resulting database for a genome-scale screening technique called GRIP-seq to globally map *cis*-regulatory elements and potentially further highlight physical connections with known TSSs. We find the GRIP-seq method robust, cost-effective and easy to implement with little or no optimization required and no theoretical limitations to its application in virtually any animal model system. The sea urchin model system we employ has a relatively complex mix of cell types and a number of thoroughly tested genes within the GRN for early development, giving us many detailed CRMs with which to validate our method. We found GRIP-seq correctly identifies known, as well as previously unknown, regulatory elements. GRIP-seq thus provides a tool for the prospective analysis of genomic regulatory elements.

GRIP-seq builds on other established methods. In particular, GRIP-seq uses ChIP-seq and 3C technology. ChIP-seq is limited to one or a few transcriptions factors at a time, which is insufficient for a complete network, especially when the identity of all active transcription factors acting at a specific time or in a specific location is unknown. The family of methods based on chromatin conformation capture (3C, 4C, 5C, and Hi-C), developed from the groundbreaking 3C method (Dekker *et al.*, 2002), are all powerful techniques, but for our specific purpose, they are either the wrong scale, only examine a single loci, or require advance knowledge of the transcription factors involved. Our method shares protocol steps with the ChIA-PET method when an anti-Pol II antibody is used in the ChIP step (Goh *et al.*, 2012). Our modifications are intended to enrich the database for active regulatory interactions and to streamline the protocol, which results in a robust and cost-efficient protocol. A table showing a side-by-side comparison of the steps in GRIP-seq and ChIA-PET is depicted in Supplemental Fig. S3.

How can we fully understand and appreciate the complete GRN of any developing embryo, distinct cancer subtype, or adult animal system physiology without comprehensive knowledge of the regulatory connections between the relevant genes? Many complementary methods at different scales will likely be needed to understand these complex regulatory systems. Due to time and resource constraints, existing gene networks have been constructed beginning with a few well-studied building blocks, with nodes and links added as experimental evidence is collected. Two

important bottlenecks to GRN analysis are the discovery of all the relevant players and the search for active regulatory modules and transcription factor binding sites. New quantitative transcriptome methods have enabled us to define the set of all factors relevant to a given network (Barsi *et al.*, 2014; Barsi, Tu, *et al.*, 2015; Fischer *et al.*, 2014; Tulin *et al.*, 2013). However, once the players are known, researchers must navigate a low-throughput bottleneck when searching for active modules regulating network genes. A comprehensive genome-wide prediction method resulting in a database of putative sites for verification will greatly speed up this process. GRIP-seq creates the necessary genome-wide regulatory database researchers require to significantly increase the speed of *cis*-regulatory analysis by drastically reducing the search space for genome sequence harboring regulatory activity. Our goals for GRIP-seq were met through development of a method that is genome-wide and applicable to all organisms. Moreover, GRIP-seq identifies sites of repression as well as activation, identifies regulation distally (and intronically) as well as proximally, and finally, often ties regulatory stretches with the TSS they govern.

We analyze the inaugural GRIP-seq database (24 hpf, mesenchyme bastula *S. purpuratus* embryos) by examining high quality mapped read counts in loci where rigorous *cis*-regulatory analysis has already been performed and published. By 24 hpf, the onset of gastrulation, the sea urchin embryo is patterned with multiple domains of differential gene expression at varying levels of resolution and overlap. This time point was therefore chosen as one that captures many critical regulatory interactions that have as a result of their important nature been previously tested by detailed experimental examination. We find good signal-to-noise ratios in every gene region we examine. The *foxa* locus shows strong signal across all four published regulatory modules. This includes a distal repressor module "F" as well as activating modules. The *foxa* locus is a good example of how researchers could have benefitted from the GRIP-seq database to increase the efficiency of screening for active regulatory regions at the time this *cis*-regulatory analysis was performed and how it can help to further complete the regulatory picture at present. In the large 20 kb region surrounding the *foxa* TSS (10 kb upstream and 10 kb downstream) there are three regions with strong GRIP-seq signal and three regions with more moderate GRIP-seq signal. Four out of these six regions, which are each close to 500 bp in size, contain regulatory activity. Such a drastic reduction in the search space for regulatory activity around a gene can increase the efficiency of *cis*-regulatory analysis. Additional analysis with peak calling software which compared each

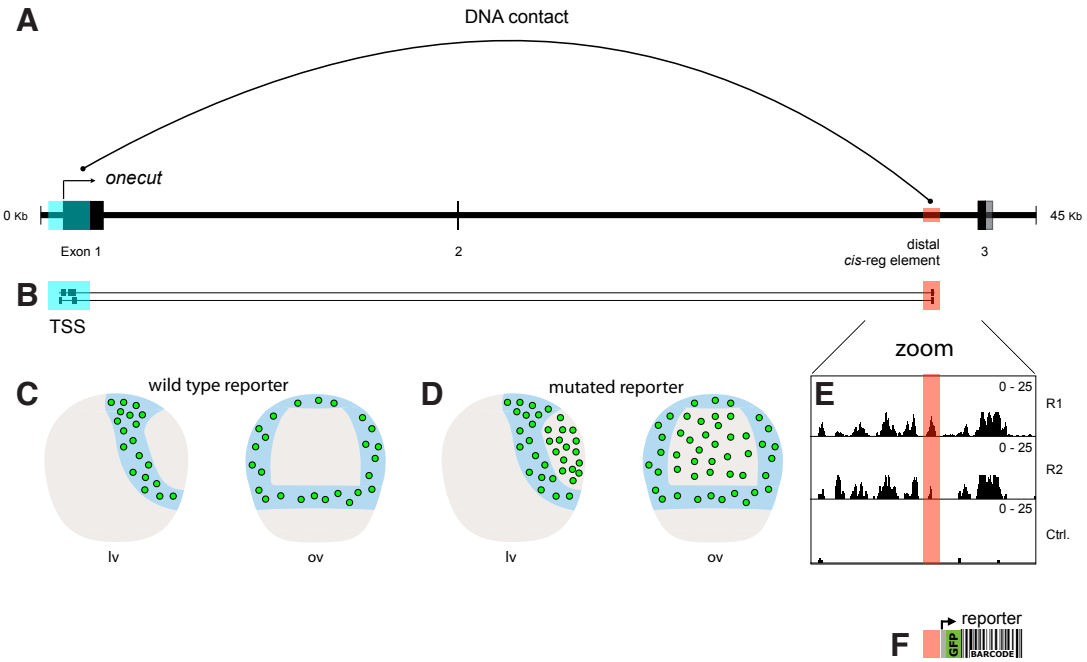


Fig. 5. GRIP-seq links a distal *cis*-regulatory element to the *onecut* TSS. *A distal* cis*-regulatory element that functions as a transcriptional repressor is shown to physically interact with the TSS of* onecut. **(A)** *Schematic of the* onecut *locus. Area highlighted in light blue indicates the TSS; grey boxes indicate untranslated regions; solid black boxes indicate exons; solid horizontal line represents 45 kb of genomic sequence; area highlighted in light red harbors regulatory sequence; an arc indicates evidence of physical contact between the DNA immediately beneath each end.* **(B)** *Paired GRIP-seq reads link* onecut*'s TSS to a distal* cis*-regulatory element found 40 kb downstream.* **(C,D)** *Schematic embryos summarize experimental evidence supporting the biological function of the distal CRE (Barsi, Davidson, 2016).* **(C)** *The zygotic expression of* onecut *at 35 hpf is confined to a circular pattern of cells from which the ciliated band will arise (shaded light blue). Transgenic embryos carrying a wild type version of the reporter shown in* **F***, are able to recapitulate this expression pattern with GFP (green circles).* **(D)** *However, when a cluster of Gsc transcription factor binding sites are mutated within the context of the same reporter, ectopic GFP expression throughout the oral face is observed, consistent with role of this CRE in repressing* onecut *expression throughout this domain.* **(E)** *GRIP-seq signal from two replicate experiments (R1 & R2) alongside their control (Ctrl.), are shown for a region immediately surrounding the distal CRE. Area highlighted in light red harbors regulatory sequence.* **(F)** *A fluorescent reporter constructed of the genomic sequence highlighted in light red, endogenous* onecut *basal promoter (grey box), GFP CDS (green box) and a DNA barcode (labeled as such) is shown to convey Gsc mediated repression onto* onecut *transcription. Abbreviations: lv, lateral view; ov, oral view.*

experimental to the control called 4,343 peaks and confirmed the pile-up views of paired reads we used for validation. Therefore, peak calling could also be used in designing *cis*-regulatory experiments by setting a peak threshold in a particular locus. In some systems, including the sea urchin, tracks are also available that depict areas of evolutionary sequence conservation, also an indication of regulation. Combining these two tracks and looking for overlap would be a useful way to incorporate two very different kinds of databases to potentially reduce the search space even further.

A limitation of using whole embryos is that quantitative analysis is not biologically significant. As a follow up to this study, a version of this method is to be used on single-cell types where the quantitative metrics will be biologically relevant. However, despite this limitation, the use of whole embryos here demonstrates the ability of this method to perform on complex cell mixtures, implying compatibility with samples from various tissues, organs, whole embryos and, by extension, tumors. Additionally, the current whole-embryo method's relevance to gene regulatory network construction is not affected by this limitation. Lastly, and most crucially, the whole-embryo approach allowed us to evaluate GRIP-seq signal at the bona fide regulatory sites rigorously established in sea urchin.

Materials and Methods

Embryo collection and fixation

Female sea urchins (*Strongylocentrotus purpuratus*) were spawned by vigorous hand shaking until 1 mL of settled egg volume was obtained. The eggs were fertilized, and diluted in 100% filtered seawater to a concentration of 1 embryo per μL of seawater. The fertilized eggs were aged for 24 hrs. at 15°C. The 24 hpf embryos were pelleted and washed twice with 10 mL of cold Phosphate Buffered Saline (PBS). One mL of embryo/PBS slurry was transferred to each of four 1.5 mL microcentrifuge tubes and fixed for 10 min in 1.11% formaldehyde at room temperature. Fixation was quenched with 65mM glycine for 5 min. Fixed embryos were pelleted and washed twice in cold PBS (1 mL) mixed with cold PMSF (10 μL). After the wash, pelleted embryos were resuspended in cold PBS (500 μL) with cold PMSF (5 μL) and pelleted again at 3,000 rpm for 1 min and stored at -80°C overnight.

Embryo lysis and genomic DNA shearing

Embryo pellets (100uL each; ~200.25 μg of DNA) were thawed on ice, resuspended in 120 μL of lysis buffer (2% Triton X-100, 1% SDS, 100 mM NaCl, 10 mM Tris-Cl at pH 8.0, 1 mM EDTA), and lysed with a 24V electric drill and a disposable plastic micropestle [Eppendorf, #022365622] in the microcentrifuge tube. The resulting lysate solution was sheared by sonication on a Covaris® S220 focused ultrasonicator using 130 μL of lysate in one microTUBE®. The Covaris® shear settings were: Peak Incident Power-105, Duty Factor-10%, 200 cycles per burst, treatment time of 130 sec, and water temperature of 7C. After sonication, the chromatin solution was transferred to a new 1.5 mL microcentrifuge tube as quickly as possible, to prevent it from sticking to the microTUBE®. The sheared chromatin was centrifuged at 13,300 rpm for 5 min, and the supernatant was removed to a new microcentrifuge tube. To check shear size, 125 μL of chromatin was incubated with proteinase K (20 mg/mL) at 65°C for 16hrs, which reverses protein cross-links. The sample was cleaned with a Zymo® column, eluted in 6.4 μL nuclease-free H_20, and 1 uL was tested on a Bioanalyzer HS DNA chip (Agilent® 2100 Bioanalyzer) to ensure an average shear size of ~500bp.

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed on the sheared chromatin, using the Millipore® Magna ChIP G kit (# 17-409). The chromatin was divided into two 53 μL samples: one for ChIP with anti-RNA Pol II antibody, the other for a negative control using Normal Mouse IgG protein. To each sample, 447 μL of kit dilution buffer and 2.25 μL of Protease Inhibitor Cocktail II was added. Antibodies were added to the samples (anti-RNA Pol II antibody (clone CTD4H8, 1uL) to the experimental samples, normal mouse IgG protein (1 uL) to the control samples). To all samples 20 μL of fully suspended protein G magnetic beads were added and samples were incubated on a nutator overnight at 4°C. Beads were pelleted on a magnetic rack and bead/antibody/chromatin complexes were washed with Low Salt Immune Complex Wash Buffer (500 μL) for 5 min and then washed with High Salt Immune Complex Wash Buffer, LiCl Immune Complex Wash Buffer, and TE Buffer, in that order. To check ChIP yield, a 5 μL aliquot of chromatin solution was removed from each sample crosslinks were reversed by incubation for 2 hrs at 62°C and 400 rpm with ChIP Elution Buffer (100μL) and 10mg/mL Proteinase K (1μL of 10 mg/mL). These yield checking samples were then cleaned with a Zymo® column, eluted in nuclease-free H_20 (10 μL), and 1 μL was analyzed on an Agilent® 2100 Bioanalyzer.

On-bead end repair, ligation, and cross-link reversal

After the last TE wash, ChIP products were immediately treated to End Repair using NEB End Repair kit, #E6050S. Beads were pelleted on the magnetic rack and nuclease free H_20 (85 μL), End Repair Buffer (10 μL), and End Repair Enzyme (5 μL) were added. Samples were incubated for 30 min at 20°C, 300 rpm, on a Fisher Scientific Isotemp® thermal shaker. Samples were removed from the shaker, beads pelleted, and the supernatant was removed. Next, H_20 (10 μL) and Quick Ligase Buffer (10 μL, #E0542S NEB) was added to samples. Samples were spun down at max speed for 10 sec and Quick Ligase Enzyme was added (1uL). Samples were vortexed gently to resuspend the beads and incubated at room temperature for 5 min, and then pelleted again. The supernatant was removed, ChIP Elution Buffer (100 μL) was added and samples were spun down. To reverse crosslinks and elute ligated products, Proteinase K (1uL) was added. Samples were vortexed gently and incubated on a thermal shaker at 62°C, 300 rpm, for 3 hrs. Next, samples were incubated at 95°C, 300 rpm, for 10 min and then cooled to room temperature. Beads were pelleted and the supernatant was collected, which now contains circularized, protein-free, DNA products.

Removal of linear DNA and amplification with rolling circle amplification

The samples were cleaned with a Zymo® spin column and eluted in 24.8 μL of H_20. Any remaining linear DNA was removed using Epicentre® Plasmid-Safe™ ATP-Dependent DNase (E3101K, 1 μL/10 units), 10x DNase buffer (3 μL), and ATP (1.2 μL), and incubated at 37°C for 30 min. Samples were purified with the Qiagen MinElute® PCR Purification Kit and then amplified by Rolling Circle Amplification using NEB® phi29 DNA Polymerase (MO269S), MCLAB® Exo-Resistant Random Primers (ERRP-100) (1 μL), Roche® PCR Nucleotide Mix (200 μM) and NEB® BSA (200 μg/mL). The hexamers and DNA were heated to 95°C for 3 min, cooled on ice for 30 sec, spun down at 13,300 rpm for 5 sec, and then added to their respective master mixes for 16 hrs at 30°C. This was followed by 10 min at 65°C to deactivate the phi29 DNA Polymerase Enzyme. Samples were cleaned with a Zymo® column, eluted in nuclease-free H_2O and stored at -20°C.

Library preparation

Sequencing libraries were constructed with an Epicentre® Nextera DNA Sample Prep Kit (# FC-121-1031), following the manufacturer's instructions. The final reaction was cleaned with a Qiagen Column, and eluted in H_2O (30μL). Samples were size selected using a 2% Pippin Prep® cassette (Sage Science), with Marker B, to select for DNA fragments between 450 bp and 550 bp. Recovered samples were cleaned with Qiagen MinElute columns and eluted in H_20 (10μL). Libraries were checked for quality by Agilent Bionanalyzer, Pico Green, and qPCR to ensure a minimum concentration of 2 nM for Illumina sequencing. The samples were pooled for sequencing in equal volumes (per-embryo weight). Libraries were sequenced on 1 lane

of an Illumina HiSeq1000 for 100bp paired ends using version 3 chemistry.

Read trimming, mapping, peak calling and visualizing

The raw reads were analyzed with FastQC (http://www.bioinformatics.babraham.ac.uk /projects /fastqc/), and then trimmed and re-paired, using quality-filtering guidelines outlined by Minoche *et al.,* 2011 (Minoche *et al.*, 2011). FastQC quality analysis was repeated to ensure filtering improved read quality. Reads were mapped onto the sea urchin genome v.3.1 using Bowtie 2, and visualized with the Broad Institute's Integrative Genomics Viewer (IGV). All raw reads and mapped read files, along with the commands used during processing and mapping are available at open access Woods Hole server (WHOAS) at https://darchive.mblwhoilibrary.org/ under doi: 10.1575/1912/7101. Data are also available in the ArrayExpress database at https://www.ebi.ac.uk/arrayexpress under accession number E-MTAB-3607 and the raw sequencing files have been transfered to the European Nucleotide Archive (ENA) for sequence searches. Peak calling was performed using MACS software, version 1.4.2 20120305, through the Galaxy Cistrome interface. The resulting peak calling.bed files for both experimental samples as compared to the control are also available in the ArrayExpress database, accession number E-MTAB-3607. The exact parameters used are included in Supplemental Fig. S4, which also includes all other README information, including Bowtie2 commands and IGV tool commands.

Reporter construct

The GFP reporter construct shown in Fig. 3 contained 1089 bp of non-coding sequence, which comprises an enhancer straddling the second, minute exon. This DNA sequence region was positioned upstream of the *onecut* basal promoter by means of fusion PCR, followed by the GFP CDS and a tailing DNA barcode. The GFP reporter construct described in Fig. 5 was identical to that published in (Barsi, Davidson, 2016).

Embryo manipulation

Microinjection of Strongylocentrotus purpuratus zygotes was performed according to well-established protocols (McMahon *et al.*, 1985). Eggs were fertilized *in situ* and zygotes injected (1 pl per zygote) with *onecut* GFP cis-regulatory reporter construct as follows: reporter construct was injected at 1 ng/μl together with 10 ng HindIII-digested genomic carrier DNA in nuclease-free water.

RNA* in situ *hybridization

Whole-mount RNA *in situ* hybridization was performed following a previously published method (Ransick, 2004). The probes used in this study were complementary to the entire CDS of *onecut* or *gfp*. Sequence information available at http://www.spbase.org:3838/quantdev (Tu *et al.*, 2014).

Accession Numbers

GRIP-seq data and peak calling files available at ArrayExpress, accession number E-MTAB-3607.

Authors' contributions

ST participated in protocol development, creating the dataset, data analysis, and drafting the manuscript. JCB performed all experiments involving the onecut *gene, participated in data analysis and drafting the manuscript. CB participated in the design of the study, performed experiments to develop the method and create the GRIP-seq dataset. JS conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.*

References

BARSI JC, DAVIDSON EH (2016). cis-Regulatory control of the initial neurogenic pattern of onecut gene expression in the sea urchin embryo. *Dev Biol* 409: 310–318.

BARSI JC, LI E, DAVIDSON EH (2015). Geometric control of ciliated band regulatory states in the sea urchin embryo. *Development* 142: 953–961.

BARSI JC, TU Q, CALESTANI C, DAVIDSON EH (2015). Genome-wide assessment of differential effector gene use in embryogenesis. *Development* 142: 3892–3901.

BARSI JC, TU Q, DAVIDSON EH (2014). General approach for *in vivo* recovery of cell type-specific effector gene sets. *Genome Res* 24: 860–868.

DANIEL B, NAGY G, NAGY L (2014). The intriguing complexities of mammalian gene regulation: how to link enhancers to regulated genes. Are we there yet? *FEBS Lett* 588: 2379–2391.

DE-LEON SB-T, DAVIDSON EH (2010). Information processing at the foxa node of the sea urchin endomesoderm specification network. *Proc Natl Acad Sci USA.* 107: 10103–10108.

DEAN FB, NELSON JR, GIESLER TL, LASKEN RS (2001). Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 11: 1095–1099.

DEKKER J, RIPPE K, DEKKER M, KLECKNER N (2002). Capturing chromosome conformation. *Science* 295: 1306–1311.

DOSTIE J, RICHMOND TA, ARNAOUT RA, SELZER RR, LEE WL, HONAN TA, RUBIO ED, KRUMM A, LAMB J, NUSBAUM C, GREEN RD, DEKKER J (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16: 1299–1309.

FENG J, LIU T, QIN B, ZHANG Y, LIU XS (2012). Identifying ChIP-seq enrichment using MACS. *Nat Protoc* 7: 1728–1740.

FISCHER AHL, MOZZHERIN D, EREN AM, LANS KD, WILSON N, COSENTINO C, SMITH J (2014). SeaBase: a multispecies transcriptomic resource and platform for gene network inference. *Integ. Comp. Biol.* 54: 250–263.

FULLWOOD MJ, LIU MH, PAN YF, LIU J, XU H, MOHAMED YB, ORLOV YL, VELKOV S, HO A, MEI PH, *et al.,* (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462: 58–64.

GOH Y, FULLWOOD MJ, POH HM, PEH SQ, ONG CT, ZHANG J, RUAN X, RUAN Y (2012). Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. *J Vis Exp.* 62: 3770.

HUTCHISON CA, SMITH HO, PFANNKOCH C, VENTER JC (2005). Cell-free cloning using phi29 DNA polymerase. *Proc Natl Acad Sci USA* 102: 17332–17336.

LEVINE M, CATTOGLIO C, TJIAN R (2014). Looping back to leap forward: transcription enters a new era. *Cell* 157: 13–25.

LI G, CAI L, CHANG H, HONG P, ZHOU Q, KULAKOVA EV, KOLCHANOV NA, RUAN Y (2014). Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* 15 Suppl 12: S11.

MCMAHON AP, FLYTZANIS CN, HOUGH-EVANS BR, KATULA KS, BRITTEN RJ, DAVIDSON EH (1985). Introduction of cloned DNA into sea urchin egg cytoplasm: Replication and persistence during embryogenesis. *Dev Biol* 108: 420–430.

MINOCHE AE, DOHM JC, HIMMELBAUER H (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 12: R112.

NAM J, SU Y-H, LEE PY, ROBERTSON AJ, COFFMAN JA, DAVIDSON EH (2007). Cis-regulatory control of the nodal gene, initiator of the sea urchin oral ectoderm gene network. *Dev Biol* 306: 860–869.

OTIM O, AMORE G, MINOKAWA T, MCCLAY DR, DAVIDSON EH (2004). SpHnf6, a transcription factor that executes multiple functions in sea urchin embryogenesis. *Dev Biol* 273: 226–243.

POUSTKA AJ, KÜHN A, RADOSAVLJEVIC V, WELLENREUTHER R, LEHRACH H, PANOPOULOU G (2004). On the origin of the chordate central nervous system: expression of onecut in the sea urchin embryo. *Evol Dev* 6: 227–236.

RANSICK A (2004). Detection of mRNA by *In situ* Hybridization and RT-PCR. In *Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental Approaches* Elsevier, pp. 601–620.

SMITH J, DAVIDSON EH (2009). Regulative recovery in the sea urchin embryo and the stabilizing role of fail-safe gene network wiring. *Proc Natl Acad Sci USA* 106: 18291–18296.

SMITH J, KRAEMER E, LIU H, THEODORIS C, DAVIDSON E (2008). A spatially dynamic cohort of regulatory genes in the endomesodermal gene network of the sea urchin embryo. *Dev Biol* 313: 863–875.

SMITH J, THEODORIS C, DAVIDSON EH (2007). A Gene Regulatory Network Subcircuit Drives a Dynamic Pattern of Gene Expression. *Science* 318: 794–797.

TU Q, CAMERON RA, DAVIDSON EH (2014). Quantitative developmental transcrip-

tomes of the sea urchin Strongylocentrotus purpuratus. *Dev Biol* 385: 160–167.

TULIN S, AGUIAR D, ISTRAIL S, SMITH J (2013). A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems. *Evodevo* 4: 16.

VAN BERKUM NL, LIEBERMAN-AIDEN E, WILLIAMS L, IMAKAEV M, GNIRKE A, MIRNY LA, DEKKER J, LANDER ES (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp.* 39: 1869.

ZHAO Z, TAVOOSIDANA G, SJÖLINDER M, GÖNDÖR A, MARIANO P, WANG S, KANDURI C, LEZCANO M, SANDHU KS, SINGH U, PANT V, TIWARI V, KURUKUTI S, OHLSSON R (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38: 1341–1347.